Third 2008 International Conference on Convergence and Hybrid Information Technology

A Decision Tree Scoring Model Based on Genetic Algorithm and K-means Algorithm

Defu Zhang, Stephen C.H. Leung, Zhimei Ye
Department of Computer Science, Xiamen University, Xiamen, 361005, China
Longtop Group Post-doctoral Research Center, Xiamen, 361005, China
Department of Management Sciences, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
dfzhang@xmu.edu.cn; mssleung@cityu.edu.hk

Abstract

Credit scoring has been regarded as a critical topic and studied extensively in the finance field. Many artificial intelligence techniques have been used to solve credit scoring. The paper is to build a classification model based on a decision tree by learning historical data. Clustering algorithm and genetic algorithm are combined to further improve the accuracy of this credit scoring model. The clustering algorithm aims at removing noise data, while the genetic algorithm is used to reduce the redundancy attribute of data. The computational results on the two real world benchmark data sets show that the presented hybrid model is efficient.

1. Introduction

The credit scoring problem is a commonly encountered decision making task in the finance field, and a typical classification problem to categorize a customer into one of the predefined classes based on a number of observed attributes related to that customer [1].

So far, credit scoring models based on data mining techniques has been applied to the finance field because information on customers' credit history has been collected and can be used to train the model. Credit scoring models can help decision-makers of banks to make more accurate decisions, and thus effectively control credit risks. Therefore, the setting up of a credit score model, which has great practical value and practical significance, has become one of the main tasks for banks [2].

Practitioners and researchers have proposed many traditional statistical methods and artificial intelligence methods for credit scoring problem [3, 4]. Currently powerful credit scoring models include neural networks [5, 6], genetic programming (GP)[7, 8], and support vector machines (SVM)[9,10]. Neural network inspired by the human brain adopts a large number of interconnected neurons and can be used to simulate non-linear relationship in complex data. It is able to deal with complex problems and can classify any customers, but its disadvantages are that the convergence speed of learning is slower and comprehensibility is not good. SVM, which is based on statistical learning theory, can deal with many classification and regression problems. Its strength lies in its ability to model non-linear data, and its high accuracy compared to other data mining techniques. However, its corresponding mathematical models are complex and lack comprehensibility. GP, which is inspired by the basic idea of Darwin's natural selection and survival principle, can generate IF-THEN rules. The main advantage of it is that it can provide intelligence classification rules for decision-makers to help them understand the contents of the data sets and make the correct decision. However, its disadvantages are that its classification capability may be bad. It may occur that a new customer does not match any rule or matches more than one rule, and thus it cannot be determined to which class it belongs. What is more, it takes longer time to build a GP model. As there is no one credit scoring model that can beat all other models for all practical problems, and that each model has its own disadvantages and merits, hybrid models which combine different credit scoring models have become more and more popular [11, 12, 13] since an improvement in accuracy of a fraction of a percent might translate into significant savings.

A decision tree is similar to a GP model, but it is relatively simpler, and therefore of practical value in many fields [14]. But decision tree is less used as a credit scoring model because its classification accuracy is easily affected by noise data and the redundancy attribute of data. Therefore, some researchers consider combining decision tree with other data mining techniques. For example, the use of decision trees

978-0-7695-3407-7/08 $25.00 © 2008 IEEE
DOI 10.1109/ICCIT.2008.110

and neural network for credit card application scoring was studied by [3], and the authors concluded that neural network and the decision tree model have a comparable level of decision accuracy. Christophe Mues et al. carried out a comparative study on decision tree application to credit scoring problems [15]. The combination of decision tree and a rough set for credit scoring problem can be found in [16], where some valuable results are obtained. In this paper, a hybrid credit scoring model is presented in which the decision tree was used as a classification model, genetic algorithm (GA) was used to remove redundant attributes, and clustering algorithm was used to remove the noise data. Computational results showed that genetic algorithm and the clustering algorithm can effectively improve the accuracy of the decision tree model. The performance of the hybrid model is verified.

2. The hybrid credit scoring model

2. 1 Decision tree

Decision tree is a very popular data mining technology in the practical field. C4.5 is a well-known induction algorithm which uses information-theoretic concepts to grow a decision tree [14]. It first grows a full tree and then retrospectively prunes it in order to avoid overfitting. C4.5 converts this tree to a set of rules which can then be further pruned.

The paper uses C4.5 to build a credit scoring model. A detailed introduction can be found in [14].

2.2 Attribute reduction based on Genetic algorithm

Generally speaking, not all attributes of data are important. There exists some redundant attributes which will affect the classification accuracy of a credit scoring model and even lead to the wrong decisions. Attribute reduction deletes some irrelevant or unimportant attributes while maintaining the attributes of classification and decision-making ability. In general, a good attribute reduction has the following characteristics: the number of attributes and reduction rules will become smaller after reduction.

There are many methods to make attribute reduction. Some methods need discretize the continuous attributes before attribute reduction. The process of discretization and attribute reduction is of independence such that it is easy to lose some important attribute information. GA can overcome the above disadvantage, so it is used to make attribute reduction [17].

GA uses binary to code the individual, the individual P is shown as follows:

$$A_1 A_2 \cdots A_n C_{11} \cdots C_{1m_1} C_{21} \cdots C_{2m_2} \cdots C_{n1} \cdots C_{nm_n}$$

where A_i denotes the i-th attribute, its value is 1 or 0, C_{ij} denotes the j-th breakpoint in the i-th attribute whether it is saved or not. m_i is the number of total breakpoints in the i-th attribute. The fitness function f can be calculated as follows:

$$\mathrm{f} = \alpha(1 - \mathrm{H}(\mathrm{P})/\log|\mathrm{U}|) + (1-\alpha)(1 - \mathrm{H}(\mathrm{D}|\mathrm{P})/\log|\mathrm{U}|)$$

Where H is the information entropy [14], α is the coefficient. D is the set of decision attribute.

The initial population depends on the generation of the candidates breakpoint set. The individual can be randomly generated. The breakpoint set of the i-th attribute A_i can be generated as follows: first obtain the r value of attribute A_i, namely $S = (V_1, V_2, \ldots V_r)$, where $V_1 < V_2 < \ldots < V_r$. Set initial breakpoint set S, and calculate the importance of each breakpoint in S, and then sort each breakpoint by their importance. At last choose the most important $k \times r$ breakpoints as a candidate breakpoint set according to a certain proportion k.

The computational method of breakpoint importance Q is as follows:

1) calculate the number L_k^j of the customers which the decision attribute value is j and the value of attribute A_i is less than V_k
2) calculate the number R_k^j of the customers which the decision attribute value is j and the value of attribute A_i is not less than V_k
3) calculate the number of the customers which the value of attribute A_i is less than V_k
4) calculate the number of the customers which the value of attribute A_i is not less than V_k
5) $Q = L_k R_k - \sum_{i=1}^{m} L_k^i R_k^i$

The choice strategy of GA is as follows: each parent individual generates k sub-individual, and then select one individual with the maximum fitness function value from k+1 individuals as the next generation individual. The crossover and mutation operation are similar to the general genetic algorithm, the only difference is that due to the coding of the individual is divided into two parts, so the corresponding operation is also divided into two parts, the breakpoint C_{ij} of each attributes is mutated and crossed respectively.

2.3 Noise elimination based on K-means

Noise data may reduce the classification accuracy of credit scoring models. We can use the clustering method to improve the prediction efficiency by removing the noise. The K-means clustering algorithm has been applied to make data preprocessing for credit scoring model and some promising results are obtained [18]. In this paper, the K-means algorithm which aims at partitioning N data items into K clusters is employed. We segment the data set into 3 clusters by the size of Euclidean distance, the cluster with the minimum customers is considered as noise data and is deleted. The detailed K-means clustering algorithm can be found at many data mining books.

As the redundancy attributes are removed by GA, using K-means to delete noise data becomes simpler. The hybrid credit scoring model is stated in Fig.1.

3. The experimental results

In order to verify the performance of the hybrid credit scoring model, the German and Australian credit data set are used to evaluate different credit scoring models. The two real world data sets are available from the UCI Repository of Machine Learning Databases [19]. The Australian data include 307 good customers and 383 bad customers. Each customer contains 6 nominal, 8 numeric attributes, and 1 decision attribute (good or bad). The German data consist of 700 good customers and 300 bad customers. It contains 20 attributes, which include 7 continuous and 13 categorical attributes.

The experimental parameters is set as follows: the probability of attribute mutation is 0.4, the number of offspring is 5, the number of the population is 5, the maximum evolution generation is 40, the threshold of the fitness value is 0.95, $\alpha = 0.1$.

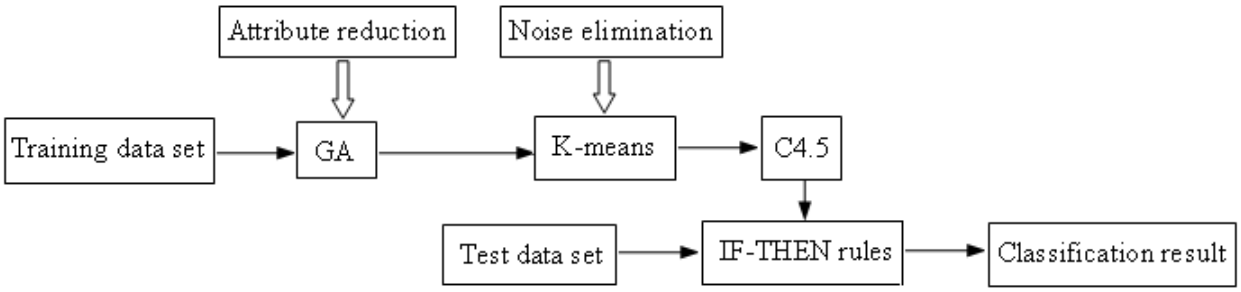

Fig.1 The hybrid credit scoring model.

Table 1 The discretization of continuous attributes in reduction attribute set

Attribute / Value	2	5	11	13	16
0	<=12.5	<=2319.5	<=1.5	<=26.5	<=3.5
1	(12.5, 15.5]	(2319.5, 2324.0]	>1.5	(26.5, 30.5]	>3.5
2	(15.5, 17.0]	(2324.0, 2326.5]		(30.5, 31.5]	
3	(17.0, 21.5]	(2326.5, 2332.0]		(31.5, 33.5]	
4	>21.5	(2332.0, 2356.0]		(33.5, 35.5]	
5		(2356.0, 2379.5]		(35.5, 37.5]	
6		(2379.5, 2392.0]		(37.5, 38.5]	
7		(2392.0, 2405.0]		(38.5, 39.5]	
8		(2405.0, 2419.5]		>39.5	
9		(2419.5, 2462.5]			
10		(2462.5, 2509.0]			
11		(2509.0, 2517.5]			
12		>2517.5			

For the Australian data, the attributes set only has two attributes after attribute reduction, the effectiveness of attribute reduction is very obvious. For the German data set, the new attribute set is (1, 2, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20) after attribute reduction. The discretization of continuous attributes in reduction attribute set is shown in Table 1.

To provide a reliable estimate and decrease the impact of data dependency in credit scoring models, *q*-fold cross validation is often used to generate random partitions of credit data sets [20]. Namely, the credit data set is divided into *q* independent groups. The credit scoring models use the (*q*-1) groups of samples as a training set and the remaining one as a test set.

Starting from the first group, the model is repeated until each group has been used as a test set once. The overall classification accuracy which is an average accuracy of all q groups is reported. In this paper, q =10. The computational results of the Australia and Germany data are shown in Table 2.

Table 2. The classification accuracy of the hybrid model for the Australian and German data(%)

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	average
Germany	78.82	77.64	78.82	76.47	81.18	64.71	78.82	80.00	83.53	77.65	77.76
Australia	86.67	88.33	90.00	95.00	93.33	86.67	91.67	81.67	90.00	90.00	89.33

Table 3 The classification accuracy of C4.5 for the Australian and German data (%)

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	average
Germany	72	72	69	79	74	60	73	76	76	80	73.1
Australia	89.85	82.61	89.86	89.86	85.51	86.96	89.86	79.71	86.96	84.06	86.52

Table 4 The classification accuracy of different scoring models for the Australian and German data (%)

	C4.5[9]	BPN	GP	SVM+GA	RSC	C4.5	The hybrid model
German	73.6	77.83	78.10	77.92	**79.63**	73.1	77.76
Australia	85.9	86.83	87	86.90	88.54	86.52	**89.33**

The computational results of C4.5 without attribute reduction and noise elimination are reported in Table 3.

We can observe Tables 2 and 3 that the classification accuracy of the hybrid model is higher than that of C4.5. It shows that attribute reduction and noise elimination are efficient and can improve classification accuracy. The decision tree size of the hybrid model for Germany and Australia is (120, 5) respectively, the decision tree size of C4.5 is (160, 40), so the decision tree of the hybrid model is simpler. In order to verify the validity of the hybrid model, we will compare the hybrid model with the C4.5, BPN, GP, SVM+GA and RSC, where the computational results of C4.5, BPN, GP, SVM+GA are directly taken from [9] in which 10-fold cross validation is adopted, RSC is taken directly from the results of the literature [16], so the comparison among the different credit scoring models is reasonable. The average cross-validation results are reported in Table 4.

We can see in Table 4 that for the Australian test data, the hybrid model outperforms C4.5, BPN, GP, SVM+ GA and RSC; For the German data sets, the hybrid model performs a little bit worse than other models. The reason may be as follows: The German data is not uniform, with the proportion of bad customers too small, being only 30 percent. The above analysis shows that the hybrid model is still effective. In addition, the results of C4.5 in [9] and C4.5 in this paper almost have no difference.

4. Conclusions

In this paper, decision tree model, which it can generate understandable rules, is used as a credit score model. In the process of establishing an initial decision tree by using a training set, the decision tree is prone to "over-fitting", thus it is necessary to use pruning techniques to improve the prediction capability for the unknown data and reduce the misjudgment rate. The redundancy attributes can affect the performance of a credit scoring model. The attribute reduction based on genetic algorithm can decrease the number of attributes and make the decision tree simper and enhance understandable. Noise data often affects the classification accuracy of credit scoring models, but K-means clustering algorithm can eliminate the affect of noise. Computational results show that GA and K-means algorithm can effectively improve the classification accuracy of a credit scoring model.

Future work includes increasing handling multiple classification attribute data, and further adjusting the parameters of genetic algorithm, or adaptively changing the calculation of the fitness value to make results more stable.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (Grant no. 60773126) and the Province Nature Science Foundation of Fujian (Grant no. A0710023) and academician start-up fund (Grant No. X01109) and 985 information technology fund (Grant No. 0000-X07204) in Xiamen University.

References

[1]Chen, M.-C., & Huang, S.-H, "Credit scoring and rejected instances reassigning through evolutionary computation techniques", *Expert Systems with Applications*, 2003, 24,pp.433-441.

[2]Treacy, W. F., & Carey, M, "Credit risk rating at large US banks", *Journal of Banking and Finance*, 2000, 24,pp.167–201.

[3]Davis RH, Edelman DB, Gammerman AJ, "Machine-learning algorithms for credit-card applications", *IMA Journal of Mathematics Applied in Business and Industry*, 1992, 4,pp.43-52.

[4]Tian-Shyug Lee, Chih-Chou Chiu, Yu-Chao Chou, Chi-Jie Lu, "Mining the Customer Credit Using Classification and Regression Tree and Multivariate Adaptive Regression Splines", *Computational Statistics and Data Analysis*, 2006, 50,pp.1113-1130.

[5]West D, "Neural network credit scoring models", *Computers & Operations Research*, 2000, 27,pp.1131-1152.

[6]Kin Keung Lai, Lean Yu, Shouyang Wang, Ligang Zhou, "Neural Network Metalearning for Credit Scoring", *ICIC (1)* 2006,pp. 403-408.

[7]Chorng-Shyong Ong, Jih-Jeng Huang, Gwo-Hshiung Tzeng, "Building credit scoring models using genetic programming", *Expert Systems with Applications*, 2005, 29(1),pp. 41-47.

[8]Qing-Shan Chen, De-Fu Zhang, Li-Jun Wei and Huo-Wang Chen, "A Modified Genetic Programming for Behavior Scoring Problem", *2007 IEEE Symposium on Computational Intelligence and Data Mining*., 2007,pp.535-539.

[9]Cheng-Lung Huang, Mu-Chen Chen, Chieh-Jen Wang, "Credit scoring with a data mining approach based on support vector machines", *Expert Systems with Applications*, 2007,33(4),pp.847-856.

[10]Jianping Li, Jingli Liu, Weixuan Xu, Yong Shi, "Support Vector Machines Approach to Credit Assessment", *International Conference on Computational Science*, LNCS, 2004, 3039,pp.892-899.

[11]Huang, Z Chen, H., Hsu, C.J., Chen, W.-H, Wu, S, "Credit Rating Analysis with Support Vector Machines and Neural Network: A Market Comparative Study", *Decision Support Systems*, 2004, 37(4),pp.543-558.

[12]De-Fu Zhang, Qing-Shan Chen, Li-jun Wei, "Building Behavior Scoring Model Using Genetic Algorithm and Support Vector Machines", Lecture Notes in Computer Science, 2007,4488,pp. 482-485.

[13] Defu Zhang, Hongyi Huang, Qing-Shan Chen and Yi Jiang, "A Comparison Study of Credit Scoring Models", *IEEE Proceedings*, ICNC 2007.

[14] Quinlan, J. R, *C4.5: Programs for machine learning*, San Francisco, CA: Morgan Kaufman, 1993.

[15] Christophe Mues, Bart Baesens, Craig M. Files, Jan Vanthienen, "Decision diagrams in machine learning: an empirical study on real-life credit-risk data", *Expert Systems with Applications*, 2004, 27,pp.257-264.

[16] XiYue Zhou, DeFu Zhang, Yi Jiang, "A New Credit Scoring Method Based on Rough Sets and Decision Tree", Lecture Notes in Computer Science, 2008,pp. 1081-1089.

[17]Fro¨ hlich, H., & Chapelle, O, "Feature selection for support vector machines by means of genetic algorithms", In: *Proceedings of the 15th IEEE international conference on tools with artificial intelligence*, Sacramento, California, USA, 2003,pp, 142-148.

[18] Michael K. Lim, So Young Sohn, "Cluster-based dynamic scoring mode", *Expert Systems with Applications*, 2007, 32,pp. 427-431.

[19] Murphy, P. M., Aha, D. W, UCI repository of machine learning databases, Department of Information and Computer Science, University of California Irvine, CA. Available from http://www.ics.uci.edu/mlearn/ MLRepository.html.

[20]Nan-Chen Hsieh, "Hybrid mining approach in the design of credit scoring models", *Expert Systems with Applications*, 2005, 28,pp.655-665.